DOCUMENT RESUME

ED 032 099 LI 001 695

By-Armstrong, Frances T.

Significant Attributes of Documents.

Georgia Inst. of Tech., Atlanta. School of Information Sciences.
Spons Agency-National Science Foundation, Washington, D.C.
Report No-GITIS-68-04
Pub Date 68
Note-15p.
EDRS Price MF-$0.25 HC-$0.85
Descriptors-*Classification, *Concept Formation, *Distinctive Features, *Documentation

The purpose of this paper is to describe a method of finding the significant attributes of documents established during the course of research on the automatic classification of documents. The problem was first approached by examining the way in which an existing hierarchical classification system classifies things. The study of biological classification lead the research into the specific study of concept formation. At that point a method was devised of applying a set of rules for forming definitions concerning the problem of concept formation. It is first necessary to have a group or set of things that are members of the concept in order to obtain the essential attributes of the concept. Given a set of things, the attributes of the members may then be listed. From the attributes listed, the essential or significant attributes must be abstracted in order to describe the concept. (Author/RM)

ED032099

LI 001695

Received in RSP 8-26-69
No. of copies 8
Grant (Contract) No. 653
FINSF
GITS 68-04

SIGNIFICANT ATTRIBUTES OF DOCUMENTS

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Frances T. Armstrong

LI 001695

Georgia Institute of Technology

School of Information Science

Work reported in this publication was performed at the School of Information Science, Georgia Institute of Technology. The primary support of this work came from the Georgia Institute of Technology and from agencies acknowledged in the report.

Unless restricted by a copyright notice, reproduction and distribution of reports on research supported by Federal funds is permitted for any purpose of the United States Government. Copies of such reports may be obtained from the Clearinghouse for Federal Scientific and Technical Information, 5285 Port Royal Road, Springfield, Virginia 22151.

* * *

The graduate School of Information Science of the Georgia Institute of Technology offers comprehensive programs of education, research and service in the information, computer and systems sciences. As part of its research activities the School operates, under a grant from the National Science Foundation, an interdisciplinary science information research center. Correspondence concerning the programs and activities of the School may be addressed to Director, School of Information Science, Georgia Institute of Technology, Atlanta, Georgia 30332.

Telephone: (404) 873-4211

GITIS-68-04

(Internal Research Memorandum)

SIGNIFICANT ATTRIBUTES OF DOCUMENTS

Frances T. Armstrong

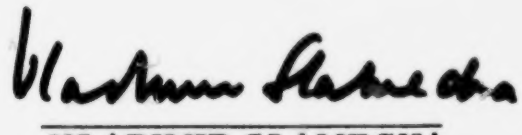
School of Information Science 1968

GEORGIA INSTITUTE OF TECHNOLOGY

Atlanta, Georgia

ACKNOWLEDGEMENT

The work reported in the paper has been sponsored in part by the National Science Foundation Grant GN-655. This assistance is gratefully acknowledged.



VLADIMIR SLAMECKA
Project Director

SIGNIFICANT ATTRIBUTES OF DOCUMENTS

The purpose of this paper is to describe a method of finding the significant attributes of documents. The method described was established during the course of research on the automatic classification of documents. The objective of that research is to develop a method (or modify an existing method) of generating an hierarchical classification system automatically.

Classification is the result of man's attempt to order knowledge: universal and fundamental classes are generated in order to understand the world. Different systems are used to classify documents, on the basis of similarities and differences in subject content. The reason for classification of documents is that it increases efficiency in locating information; therefore, a classification system is a method for grouping material so that related documents are together.

It should be pointed out that the classification of materials such as books or documents is considered an art, whereas the classification of things is in nature itself, and is the true order of the sciences. The order of the sciences therefore is the foundation for the classification of material, but modifications may be made as determined by the complexity of the material as well as by the reason for the classification, which is to facilitate use of the material. It has been noted, though, that the closer a classification system is to the order of the sciences, the better the system will be and the longer it will remain valid.

To classify documents it is first necessary to obtain the characteristics or attributes which will describe each document. Attributes have been distinguished in a number of different ways:

(1) Attributes may be thought of as either <u>essential</u> or <u>accidental</u> attributes. Essential attributes give the primary nature of a thing — that without which the thing could not be itself. In contrast, accidental attributes can be changed without affecting the primary nature of the thing. It should be pointed out that attributes essential to a particular thing are not necessarily essential to some other thing, and that attributes essential to a subclass are not necessarily essential to the larger class.

(2) Attributes may be thought of as either primary or secondary attributes. Primary attributes are attributes which exist in an object independent of an observer. Secondary attributes exist through the senses of an observer.

(3) Attributes may be thought of as being certain kinds of attributes and also as having different degrees. Each individual attribute is a kind of attribute. An attribute that does not vary or have any variable relations expresses only a kind of attribute. An attribute which varies has a difference of degree and expresses more or less of the quality.

This paper is concerned explicitly with the problem of identifying the essential attributes or essential characteristics of a set of documents. The problem was first approached by examining the way in which an existing hierarchical classification system classifies things; this was done to try to establish how the essential attributes are known. The system chosen for study was biological classification, or taxonomy for animals. The study of that system lead our research into the specific study of concept formation. At this point, we devised a method of applying a set of rules for forming definitions to the problem of concept formation.

Biological classification is a natural classification. A natural classification is based on what are called the essential attributes of the things to be classified. But what are the essential attributes? It has been stated that the essential attributes are associated, universally or in a high percentage of all cases, with other attributes of which they are logically independent. (1)

Animal taxonomists maintain that to describe an animal one must take into consideration its structure, distribution, genetics, mode of life, and physiology — in other words, all its aspects. Attempting to describe something by using only a single attribute not only will result in the grouping together of unrelated forms, but will in some cases be impossible, since there may not be one attribute to rely on. It is therefore necessary, in grouping similar animals together, to take into account all features, and to look for general resemblances and general differences to form a concept.

Another consideration is the weighting of the attributes. All attributes must be taken into account, but not all are of equal importance. With animals, some attributes are adapted to a mode of life, and the importance of the attributes must be reduced for classification.

One way of establishing the importance of an attribute in a group is to test its constancy within subgroups constructed by considering all the other attributes. In some groups a certain attribute may be extremely important, while in other groups the same attribute may be of little consequence. The importance of an attribute within a group depends on how extensively its occurrence within that group is correlated with all other attributes; therefore, the essential attributes of a group are those which, after consideration of all the attributes of a group, are found to be most useful in defining the group. (2)

The results of this study indicate that to find the essential attributes or essential characteristics for a set of objects, it is necessary to have a knowledge of the background of the objects, and to consider <u>all</u> attributes. However, it will be found that some attributes are of no importance to the classification system, and that the important attributes are not given equal weight over the entire system. An attribute may be of extreme importance in describing one concept of the system but of little value in describing another concept.

It would seem that it would be easier to classify animals than documents, since animals are objects and, as objects, possess attributes which are available to the senses; whereas the attributes of documents are words or word phrases and can be dealt with only by dealing with the language. And, in fact, the automatic classification systems studied classify documents on the basis of the words contained in them, since the ideas in the documents are expressed in words or word phrases. This means that in order to classify a document, one must form a concept using the words and word phrases as the attributes.

Concept formation involves a common identifying response that is associated with items that are not completely identical. Three types of concepts can be considered:

(1) Conjunctive concept. The members of the concept have at least one common attribute or one common group of attributes.

(2) Disjunctive concept. The members of the concept do not have one common attribute or one common group of attributes, but do have at least one attribute of a group of attributes.

(3) Relational concept. The members of the concept do not have one common attribute or one attribute of a group of attributes, but the members of the concept show a certain relationship or follow some set of rules.

To form a concept given a set of documents and the attributes pertaining to the documents, the conjunctive concept is used. The attribute itself may be a single aspect, a group of aspects that are joined conjunctively, a group of aspects that are joined disjunctively, or a relation.

Cassirer (3) has written extensively on concept formation or class formation in language, and his thoughts seem applicable. He states:

> The problem of concept formation marks the point of closest contact between logic and the philosophy of language; at this point they seem to fuse into an inseparable unit. For all logical analyses of concepts seem eventually to lead to the study of words and names. (4)

> Traditional logic tells us that the concept arises "through abstraction": it instructs us to form a concept by comparing similar things or percepts and abstracting their "common characteristics." That the contents of comparison have specific "characteristics," that they possess qualitative properties according to which we can divide them into classes, genera, species is usually taken as a self-evident premise, requiring no special mention. And yet this seemingly self-evident premise embodies one of the most difficult problems of concept formation. (5)

> In the usual logical view, the concept is born only when the signification of the word is sharply delineated and unambigously fixed through certain intellectual operations particularly through "definition" according to genus proximum and differentia specifia. But to penetrate to the ultimate source of the concept our thinking must go back to a deeper stratum, must seek those factors of synthesis and analysis,

> which are at work in the process of word formation itself, and which are decisive for the ordering of all our representations according to specific linguistic classifications. (6)
>
> Before any contents can be compared with one another and ordered into classes according to the degree of their similarity, they themselves must be defined as contents. (7)

To understand linguistic concept formation one must see how language progresses from a qualificative to a generalizing view; from the concrete to the universal. This can be done by comparing the concepts of advanced languages with the concepts of primitive languages.

> The languages of primitive peoples designate every thing, every process and activity, with the most intuitive concretion; they strive to express as plainly as possible all the distinguishing attributes of a thing, all the concrete details of an occurrence, every modification and shading of an action. In this respect they possess a richness which our advanced languages cannot even begin to approach. (8)
>
> ... before language can create specific class designations and "generic concepts," it concentrates on the designation of "varieties." (9)

The naming of every variety may also occur in highly developed languages. It is felt that this individualizing occurs because we sharply individualize that which has more meaning, importance, or interest to us. It also seems that we individualize what is <u>new</u> to a language, even if the language is advanced; and that it takes a certain amount of time to begin generalizing and forming concepts of the new entries. In other words, one must stand back and get the over-all picture.

> The genuine concept does not disregard the peculiarites and particularities which it holds under it, but seeks to show the <u>necessity</u> of the occurrence and connection of just these particularities. What it gives is a universal <u>rule</u> for the connection of the particulars themselves. (10)

It is first necessary to have a group or set of things that are members of the concept in order to obtain the essential attributes of the concept. Given a set of things, the attributes of the members may then be listed. From the attributes listed, the essential or significant attributes must be abstracted in order to describe the concept.

It is stated that the definition of a general term is a description of all members of a class or a concept; this description has a special purpose — to give just those attributes which will mark out or delimit that class from other classes. Since defining a term and forming a concept are closely related, the method devised here to obtain the significant attributes of a concept is a method used for defining terms — definition by genus and differentia.

Before proceeding, however, the terms "extension" and "intension" need to be defined. "Extension" is a synonym for "denotation," and "intension" is a synonym for "connotation." The extension of a term or a concept is the sum total of all the members (or documents, in our case) to which the term or concept refers. The intension of a term or a concept is the set of attributes which the members of a concept must possess to be within that concept. Therefore, by the intension of a concept we mean the essential attributes of the concept. It should also be pointed out that the extension and the intension of a concept vary inversely: the fewer the members of a concept, the greater the number of common attributes. However, this inverse variation depends also upon the degree of difference between members of the concept.

A definition by genus and differentia first places the term to be defined in a larger class and then eliminates the nonrelevant subclasses of this larger class by stating the essential intensional attributes which are possessed only by the class being defined. The genus marks off and focuses attention upon a large general area, whereas the differentia, as a statement of the essential intension, delimits. For a definition to delimit an extension as precisely as possible, the differentia must state both the <u>necessary</u> and the <u>sufficient conditions</u> which a thing must possess to belong to the class in question. (11)

The following is the procedure for formulating a definition — or, in our case, a concept. It is also, of course, a procedure for obtaining the significant attributes:

1. Obtain a set of examples which are members of the concept or class in question. These examples should be varied and particular, and should cover the entire area and include the borderline cases which seem to be part of the concept. The attributes of this set should then be listed. (An attribute, here, is not necessarily a single aspect, but may be a complex set of aspects or a relation.)

2. Obtain a set of examples which are <u>not</u> members of the concept or class in question; these examples should include the borderline cases which seem <u>not</u> to be part of the concept. The attributes of this set should also be listed.

3. The appropriate genus for the concept must contain all the members of the concept and be capable of containing members that are not members of the concept.

4. The appropriate differentia must state the <u>necessary</u> and <u>sufficient conditions</u> for membership. From the examples, select the parts, qualities, relations and functions which are the essential and delimiting aspects. These should be the essential or significant attributes and, when taken together, should pertain to all members of the concept being described (and not to nonmembers). Here, parts are separate units: qualities are features or unitary aspects; relations are connections between related units or aspects; and functions involve action or changing aspects.

5. Obtain the significant attributes of the concept by comparing the attributes of the two sets of examples. The significant attributes are positive and negative. The positive attributes must be a part of the set which are members of the concept and the negative attributes must be a part of the set which are not

members of the concept (and not part of the first set).

The next phase of research on significant attributes is to apply the procedure outlined herein on a set of simple data, and then evaluate the results.

List of References

1. Carl G. Hempel, Fundamentals of Concept Formation In Empirical Science, International Encyclopedia of Unified Science, Vol. II, No. 7, University of Chicago Press, Chicago, 1952, p. 52.

2. A. J. Cain, Animal Species and Their Evolution, Harper and Brothers, New York, 1960, pp. 11-26.

3. Ernst Cassirer, The Philosophy of Symbolic Forms, Vol. I: Language, Yale University Press, New Haven, Connecticut, 1966.

4. Cassirer, p. 278

5. Cassirer, p. 279

6. Cassirer, p. 280

7. Cassirer, p. 280

8. Cassirer, p. 289

9. Cassirer, p. 290

10. Ernst Cassirer, "On the Theory of the Formation of Concepts," Pattern Recognition, L. Uhr (ed), Wiley and Sons, New York, 1966, p. 30.

11. Hubert G. Alexander, Language and Thinking - A Philosophical Introduction, D. Van Nostrand Company, Inc., New York, 1967.

Bibliography

Alexander, Hubert G., Language and Thinking - A Philosophical Introduction, D. Van Nostrand Company, Inc., New York, 1967.

Atherton, Pauline (ed), Classification Research, Proceedings of the Second International Study Conference, 1964, Munksgaard, Copenhagen, 1965.

Cain, A. J., Animal Species and Their Evolution, Harper and Brothers, New York, 1960.

Cassirer, Ernst, "On the Theory of the Formation of Concepts," Pattern Recognition, L. Uhr (ed), Wiley and Sons, New York, 1966.

Cassirer, Ernst, The Philosophy of Symbolic Forms, Vol. 1: Language, Yale University Press, New Haven, Connecticut, 1966.

Haygood, R. C. and Bourne, L. E., "Attribute-and-Rule-Learning Aspects of Conceptual Behavior," Psychol. Rev., 72, No. 3, 1965, p. 175-195.

Hempel, Carl G., Fundamentals of Concept Formation In Empirical Science, International Encyclopedia of Unified Science, Vol. II, No. 7, University of Chicago Press, Chicago, 1952.

Hunt, Earl B., Concept Learning, An Information Processing Problem, John Wiley and Sons, Inc., New York, 1962.

Hunt, Earl B., Marin, Janet and Stone, Philip J., Experiments In Induction, Academic Press, New York and London, 1966.

Kendler, Tracy S., "Concept Formation," Annu. Rev. Psychol. 13, 1961, pp. 447-472.

Klausmeier, Herbert J. and Harris, Chester W. (eds), Analyses of Concept Learning, Academic Press, New York and London, 1966.

Manis, M., Cognitive Processes, Wadsworth Publishing Company, Inc., Belmont, Calif., 1966.

Pikas, Anatol, Abstraction and Concept Formation, Harvard University Press, Cambridge, Mass., 1966.

Richardson, Ernest Cushing, Classification, Theoretical and Practical, The Shoe Strong Press, Inc., Hameden, Connecticut, 1964.

Trachenberg, A., "Automatic Document Classification Using Information Theoretical Methods," Automation and Scientific Communication, Part 2, Luhn, H. P. (ed), American Documentation Institute, Washington, D. C., 1962.

Vickery, B. C., Classification and Indexing In Science, Butterworths Scientific Publications, London, 1958.

Ward, J. H., Jr., and Hook, Marion E., "Application of an Hierarchical Grouping Procedure to a Problem of Grouping Profiles," Educ. and Psych. Measurement, 23 (Spring 1963) pp. 69-83.